분석 비용을 줄여주는 다중 서열 수집과 번역을 위한 생물정보학 도구

A Labor-Saving Bioinformatics Tool for Multiple Sequence Collection and Translation

이승희[1], 이혜리[1], 이건명[1], 이찬희[2]

[1] 충북대학교 전기전자컴퓨터공학부
[2] 충북대학교 생명과학부
E-mail: shlee@aicore.cbnu.ac.kr

요 약

많은 생물학적 데이터베이스와 도구들이 네트워크 상에서 이용가능하다. 데이터베이스와 도구를 효과적으로 활용하면, 비용을 줄이면서 우수한 품질의 분석결과를 얻을 수 있다. 이 논문에서는 서열분석시 관련된 서열을 자동으로 수집하여, 아미노산 서열로 변환하는 도구에서 대해서 소개한다. 개발된 도구는 필요한 서열을 주어진 질의를 기반으로 하나의 DNA 서열 정보와 관련된 서열을 검색하도록 하고, 분석자가 관심있는 항목을 쉽게 선택하게 하여, 이것을 아미노산 서열로 번역하고, 찾은 ORF를 기반으로 유사한 것을 추천하고, 번역된 ORF 서열과 어울리는 관련된 모든 정보를 검색하는 분석 과정을 자동화한 것이다.

Key Words : 생물정보학, 서열 변환, DNA 서열, 아미노산 서열, ORF

1. 서 론

분자 생물학자들과 생물정보학자들의 개방된 사고방식은 분석자들을 위해 네트워크 상에서 무료로 이용 가능한 많은 유용한 데이터베이스와 분석 도구들을 만들게 하였다. 생물정보학 분석은 일반적으로 다양한 데이터베이스와 도구의 사용을 요구한다. 개개의 도구와 데이터베이스들은 능률적이지만, 다양한 데이터베이스와 도구를 함께 사용하는 것이 때로는 번거롭다. 다른 도구들을 사용하기 위해서 효과적으로 구성된 도구와 데이터베이스들은 분석을 하는 수고를 줄이는데 큰 기여를 할 수 있다.

정렬하는 기술들과 효과 있는 계산적 도구들의 높은 처리량으로 인해 많은 유기체를 구성하는 염색체들은 생물학적 데이터베이스로 잘 정리되고 보관되었다. 많은 연구자들은 유전자 기능의 식별, 유전자 제어 네트워크와 같은 상호작용 네트워크의 설계, 신호 변환 네트워크와 대사 경로 네트워크와 같은 post-genome 연구에 주의를 기울였다.

하나의 서열은 어떤 생물학적 처리로부터 획득되는데, 생물학적 분석자는 다른 관심 있는 서열에 대해서 DNA 기반 수준 또는 아미노산 수준에서 이것의 유사성과 차이를 평가하여 구조적, 기능적, 그리고 진화적 관계의 추론을 시도한다[1]. 생물학적 분석자들은 일반적으로 NCBI[2]와 같은 알려진 생물학적 데이터베이스로부터 존재하는 서열들과 대응하는 주석 정보를 수집하고, 웹 애플리케이션과 독립 애플리케이션으로 지정된 분석을 한다.

다양한 생물정보학 도구들은 최적화되거나 접근된 해를 찾고, 복잡한 연산을 수행하고, 효율적으로 분석자들에 의해 수동으로 완료되는 분석을 하기 위해서 개발되었다.

본 연구는 알려진 데이터로부터 다중 DNA 서열들을 얻고, 이것을 아미노산 서열들로 변환하기 위해서 분석자가 필요한 분석 상황이 고려되었다. 이 작업을 하기 위해서 먼저, GenBank(GB) 등록 번호나 Geninfo(GI) 식별자를 NCBI GenBank 데이터베이스[3]에 입력하고, 검색된 결과로부터 DNA 서열 부분을 추출하고, 파일에 추출된 결과를 저장한다. 수집된 서열들을 수작업이나 ExPAsy의 Translate[4]와 같은 번역 프로그램을 사용하여 아미노산 서열들로 하나씩 번역하고, 요구되는 형식으로 이들을 파일에 저장한다.

본 논문에서는 위에서 설명한 모든 과정을 한 번에 수행하기 위해서 설계되고 구현된 지능적 도구를 보여준다. 본 논문은 다음과 같이 구성되었다. 2절에서는 DNA 서열 번역 도구들과 관련된 연구에 대해서 살펴보고, 3절에서는 시스템 구조와 이것의 기능성에 대해서 제안된 접근 방법과 개발된 시스템을 보여준다. 4절은 생물학적 분석에서 제안된 도구를 사용하여 얼마나 효과적인가를 논의한다. 5절은 결론과 향후 연구 과제에 대해서 다룬다.

이 논문은 2007년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(지방연구중심대학육성사업/충북BIT연구중심대학육성사업단)

2. 관련 연구

생물학적 정보를 얻기 위해서 사용되는 대표적인 데이터베이스에는 GenBank, EMBL[5]의 UniProtKB/Swiss-Prot, PIR-International 등이 있다. 이들은 질의 인터페이스를 갖추고 있고, 웹 페이지나 ASN.1, XML, FASTA와 같은 다양한 표준 형식으로 검색된 결과를 제공한다.

이러한 데이터베이스에서 각 레코드는 서열, 주석, 유전체, 참고문헌 등을 포함한 다양한 필드를 구성한다. 데이터베이스들은 같은 개체들에 대해서 부분적으로 겹쳐진 레코드들을 가지고 있다. 생물학적 데이터베이스에서 각 레코드는 등록 번호라고 부르는 고유의 식별자를 가지고 있는데, 이것과 레코드들이 데이터베이스 사이에서 연결된다.

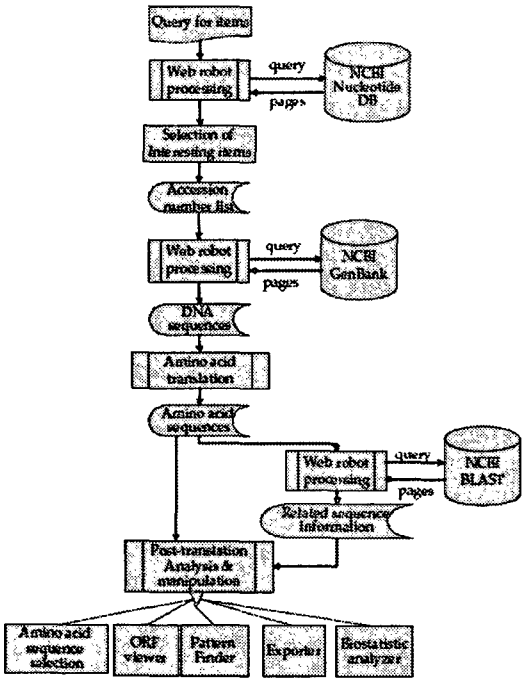
GenBank는 일반적으로 새롭게 식별된 서열을 등록하기 위해 사용되는 데이터베이스이다. 이것은 DNA와 단백질 서열, 유전자 지도, 분자 설계 데이터베이스, 주석 문서에 대한 데이터를 가지고 있다. GenBank를 위해서 NCBI는 Entrez라고 하는 웹 질의 인터페이스 프로그램을 제공하고, 이것의 기능 중 하나는 GenBank 등록 번호나 Geninfo 식별자로 관련 정보와 함께 DNA 서열들을 검색하는 것이다 [2]. Entrez는 사용자가 DNA 서열을 추출하기 위해서 텍스트, XML 등 다양한 형식으로 검색된 결과를 나타낼 수 있다.

ExPAsy(Expert Protein Analysis System)는 Translate라고 하는 프로그램을 제공하는데, 이것은 주어진 서열을 대응하는 아미노산 서열로 번역한다 [4]. ExPAsy는 SIB에 의해 운영되는 단백질체학과 관련된 정보를 제공하기 위한 시스템이다. Translate는 하나의 DNA 서열을 한번에 변환한다. 그러므로 다중 DNA 서열들을 다루기 위해서 사용자들은 번역되기 위한 서열들의 수만큼 많은 시간의 번거로운 잘라내기와 붙여넣기 작업을 해야 한다.

사용자가 관심 있는 다중 DNA 서열을 수집하고, 이들을 대응하는 아미노산 서열들로 번역할 때, 다음과 같은 과정을 수행한다. NCBI Entrez로 등록 번호에 의한 DNA 서열을 검색하고, 검색된 결과에서 수작업으로 DNA 서열 부분을 추출하여 FASTA 형식으로 번역한다. 그리고 ExPAsy의 Translate에 DNA 서열들을 제공하여 하나씩 번역된 서열들을 파일로 잘라내고 붙여넣기를 한다. 이와 같은 과정은 시간 낭비이고, 번거로운 작업이다. 이러한 관찰로 이 본문에서는 최소화된 사용자의 관여로서 모든 과정을 다룰 수 있는 시스템을 개발하였다.

3. 개발된 시스템

자동으로 다중 DNA 서열을 획득하고 번역을 하는 시스템은 DNA 서열 획득 모듈, 번역 모듈, 그리고 Pattern Finder, Biostatistics Viewer, Exporter와 같은 부가적인 도구들로 구성하여 개발하였다. DNA 서열 획득 모듈은 웹 로봇 기술을 사용하여 구현되었는데, 이것의 역할은 DNA 서열들을 NCBI GenBank로부터 수집하는 것이다. 번역 모듈은 수집된 DNA 서열들을 아미노산 서열들로 번역하는 것을 담당하는데, 여기에서 가능한 reading frame의 위치와 방향에 따라서 각 DNA 서열에 대해 6개의 아미노산 서열들이 생성되고, 가장 적절한 하나가 추천된다. Pattern Finder는 선택된 DNA 서열에서 특별한 부분 서열을 검색을 한다. Biostatistics Viewer는 서열들에 대해서 기본적인 통계 정보를 보여준다. Exporter는 처리된 서열들을 파일로 출력한다.



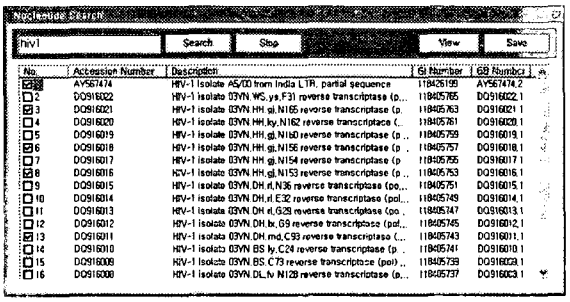
[그림 1] 개발된 시스템의 작업 흐름

사용자가 이 시스템을 사용할 때, 등록 번호 리스트를 사용하여 NCBI GenBank로부터 DNA 서열들의 수집을 요청하거나 번역 모듈을 위해서 직접 DNA 서열들을 제공할 수도 있다. 수집되기 위한 DNA 등록 번호들은 파일 안에 목록으로 저장되어 있고, 파일은 시스템에 입력 파일로서 사용된다. 사용자에 의해 수집된 DNA 서열들을 저장하기 위한 파일은 FASTA 형식으로 편집이 가능하고, 시스템은 아미노산 서열들로 번역하기 위해서 이 파일을 사용하기 위한 기능을 제공한다. 각 DNA 서열들에 대해서 시스템은 6가지 가능한 번역 방법을 가지고 있고, 첫 번째 위치에서 가장 유사한 하나를 추천한다. 그리고 사용자 인터페이스를 통해서 사용자가 그것을 확인하거나 번역된 서열들로부터 다른 것을 선택하는 것이 가능하다. 확인된 아미노산 서열들은 FASTA나 XML 형식의 파일로 출력된다. [그림 1]은 다중 DNA 서열 번역을 위한 개발된 시스템의 작업을 보여준다.

3.1 정보 수집을 위한 질의 로봇

질의 로봇은 분석자의 질의를 검색하는데, 질의를 NCBI Entrez로 보내서 응답을 받고, 관심 있는 정보를 얻기 위한 응답 결과를 분석한다. [그림 2]

와 같이 질의 키워드와 관련된 검색 정보를 보여주고, 분석자들이 NCBI nucleotide 데이터베이스로부터 DNA 서열이 검색된 결과 중에서 항목을 선택할 수 있다.



[그림 2] 질의 로봇 인터페이스

3.2 DNA 서열 수집을 위한 웹 로봇

NCBI GenBank로부터 관심 있는 다중 DNA 서열들을 수집하기 위해서 NCBI Entrez와 상호작용하는 웹 로봇을 개발하였다. 웹 로봇은 먼저 DNA 등록 번호들을 포함한 파일을 업로드하고, 등록 번호에 대응하는 DNA 서열들을 GenBank에서 하나씩 검색한다. 그리고 수집된 서열들은 내부 데이터 구조로 유지되고, 트리 뷰 디렉토리를 통해서 보여준다.

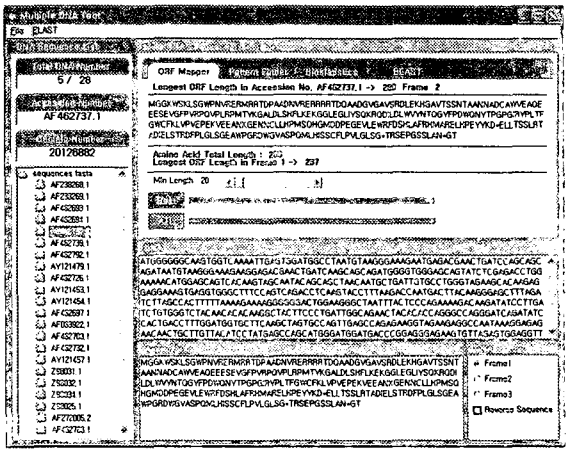
각 등록 번호에 대해서 웹 로봇은 Entrez CGI 프로그램에 보내기 위한 질의를 생성한다. CGI프로그램은 각 질의에 대한 응답으로 GenBank nucleotide 데이터베이스에서 레코드에 대응하는 것을 검색하고, 검색된 결과를 돌려보내 준다. 그리고 로봇은 검색된 결과를 분석하여 DNA 서열을 추출한다. 수집된 DNA 서열들은 FASTA 형식 데이터 구조로 유지되고, 다음 번역 작업에 사용된다. 웹 로봇은 서열들을 수집는데 있어서 번거롭고 시간을 낭비하는 Entrez 시스템과의 수동적인 상호작용을 회피하여 분석자에게 도움을 준다.

3.3 DNA 서열 변환

DNA 서열은 아데닌(A), 티민(T), 구아닌(G), 시토신(C)을 나타내는 문자들로 구성되고, 유전 정보를 부호화한다. 각 연속되는 3문자들은 코돈이라고 부르는데, 이것은 하나의 아미노산으로 부호화할 수 있다. 코돈은 20개의 아미노산에 대해서 64개의 표현 방법이 있다. 각 코돈은 아미노산 서열에서 한 문자와 대응된다[8].

DNA 서열에 대해서 그들은 각각 5′에서 3′, 3′에서 5′ 방향으로 3가지 가능한 reading frame이 존재한다. reading frame은 DNA 서열을 아미노산 서열로 번역하기 위한 읽는 방식인데, 이것은 DNA 서열 중 어느 부분부터 읽을 것인가에 따라서 다른 아미노산 서열을 생성하게 된다. 개발된 시스템은 6개의 아미노산 서열을 각 reading frame에 따라서 생성 한다. [그림 3]과 같이 DNA 서열이 선택될 때, 사용자 인터페이스는 아미노산 서열들을 보여주고 6가지 출력된 서열들 중에서 하나를 선택할 수 있다.

코돈의 서열인 각 reading frame에서 개시 코돈(ATG)으로 시작하여 종결 코돈(TAA, TGA, TAG)으로 끝나는 부분을 ORF(Open Reading Frame)라고 한다. 하나의 ORF는 유전자를 부호화하고 있는 곳을 코드화한 영역인데, 이것은 단백질로 표현될 때 사용된다. 생물학에서 ORF는 잠재적인 단백질 구성 정보를 부호화한 서열을 결정하거나 코드화된 단백질의 크기나 분자량을 평가하기 위한 척도로서 사용된다[9]. 의미 있는 ORF는 일반적으로 가장 긴 reading frame이다.



[그림 3] 개발된 다중 DNA 수집과 번역을 위한 도구

가장 긴 ORF를 찾기 위해서 개발된 시스템은 개시 코돈(M)부터 종결 코돈(*)까지 가능한 조합을 사용한다. 분석자에 의해 제어될 수 있는 최소 ORF 길이의 임계 값으로서, 임계 값보다 큰 크기의 가장 긴 ORF가 선택된다.

DNA 서열에 대해서 가장 유사한 아미노산 서열이라는 것을 언급하기 위해서 개발된 시스템은 6개의 가능한 방법으로 번역된 아미노산 서열들 중에서 가장 긴 ORF를 사용하여 결정한다. 가장 유사한 서열은 등록 번호 리스트 윈도우에서 등록번호가 선택될 때, 체크 박스가 체크된 상태에 따라서 Amino Acid Viewer에 보여주게 된다. 이러한 추천은 적은 노력으로 번역된 아미노산 서열을 확인할 수 있도록 분석자에게 도움을 주고, 그들은 추천된 것 이외의 다른 번역을 선택하여 볼 수 있다. 확인된 서열은 차후에 파일로 출력된다.

3.4 BLAST와 통신하기 위한 웹 로봇

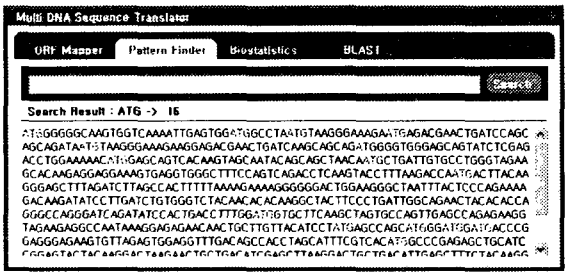
저장된 서열들은 추후에 BLAST(The Basic Local Alignment Search Tool)와 같은 서열 유사성 기반 검색 서비스를 위한 질의 서열로 사용된다.

3.5 ORF Mapper

ORF Mapper는 선택된 아미노산 서열에 대해서 추천된 ORF를 보여주는 윈도우이다. 이 윈도우에서 가능한 ORF의 위치는 막대기 그래프로 표현된다. ORF의 최소 길이에 대한 임계 값은 분석자에 의해서 슬라이더바 인터페이스를 사용하여 제어될 수 있다. 윈도우는 아미노산 서열 길이와 가장 긴 ORF 길이에 대한 단순 통계를 출력한다.

3.6 Pattern Finder

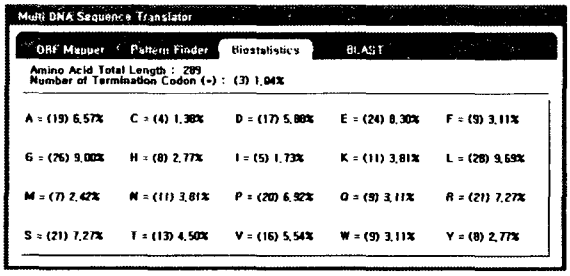
Pattern Finder는 선택된 reading frame을 기반으로 패턴에 대한 검색을 도와주는 도구이다. [그림 4]와 같이 윈도우에서 DNA 서열 중 일치된 부분은 쉬운 추적을 위해 다른 색으로 바뀌게 되고, 질의 패턴이 나타난 횟수도 보여준다. 분석에서 자주 검색되는 패턴은 시작 코돈, 종결 코돈, att site 등이다.



[그림 4] Pattern Finder

3.7 Biostatistics Viewer

Biostatistics Viewer 윈도우는 번역된 서열에 대한 통계적 정보를 보여준다. 개발된 시스템의 현재 버전에서는 선택된 아미노산 서열에 대해서 아미노산의 구성과 분포에 대한 정보를 제공한다. [그림 5]는 Biostatistics Viewer를 보여준다.

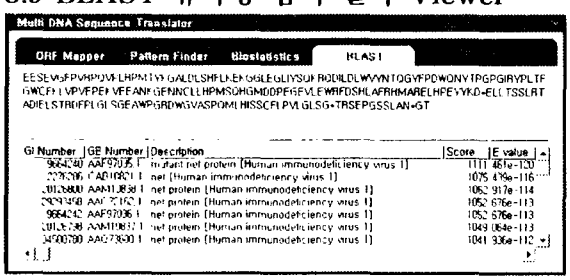


[그림 5] Biostatistics Viewer

3.8 Exporter

차후에 다른 처리를 하기 위해서 exporter는 GenBank로부터 수집된 DNA 서열들과 번역된 아미노산 서열들을 파일로 출력한다. 이것은 Genbank로부터 수집된 DNA 서열들과 각 DNA 서열에 대해서 6가지 가능한 방법으로 번역된 아미노산 서열들을 파일로 출력한다. 파일은 FASTA 형식이나 XML 형식으로 저장될 수 있다.

3.9 BLAST 유사성 검색 결과 Viewer



[그림 6] BLAST 유사성 검색 결과 Viewer

BLAST 유사성 검색 결과 Viewer는 선택된 DNA 서열에 대응하는 번역된 아미노산 서열을 NCBI BLAST 데이터베이스에 저장된 서열들과 비교하여 가장 유사성이 높은 순서로 결과를 보여준다. [그림 6]은 BLAST 유사성 검색 결과 Viewer를 보여주고 있다.

3.10 구현

이 시스템은 윈도우 환경을 위해 Microsoft Visual Basic 컴포넌트를 사용한 독립 애플리케이션으로 구현되었다. 애플리케이션은 NCBI GenBank에 접근하기 위해서 인터넷에 접속된 컴퓨터에 설치되어야 한다.

4. 제안된 도구의 효과 분석

개발된 도구의 목적은 정보 획득, 서열 수집과 번역에서 분석자의 간섭을 줄이기 위한 것이다. 분자생물학자들이 한국형 HIV 바이러스 B형을 찾는다고 가정할 때, 분석자는 먼저 NCBI nucleotide 데이터베이스를 검색하여 HIV B형에 대한 정보를 수집하고, 비교 작업에 대한 분석을 위한 항목을 결정한다. 각 선택된 항목은 페이지에 따라서 검색하고, GI number, 주석 등과 같이 다른 관련된 데이터에 따라서 수작업으로 DNA 서열 데이터를 추출한다. 각 항목을 사용하기 위해서 이 작업 이후로 약 15초 후에 분석자는 찾기, 잘라내기, 붙여넣기 작업과 같은 번거로운 작업을 몇 분 동안 해야 한다. 개발된 도구는 분석자의 이익을 위해 이러한 작업을 한다.

전통적인 분석 방법으로 DNA 서열을 아미노산 서열에 대응하는 것으로 번역하기 위해서 ExPAsy의 Translate가 사용될 수 있다. Translate를 사용할 때 분석자는 DNA 서열을 복사해야 하고, Translate 인터페이스에 붙여넣기를 해야 한다. 번역된 결과는 수작업을 통해 파일로 복사된다. 이와 같은 번역 작업은 각 DNA 서열에 대해서 하나씩 이루어진다. 이러한 수작업은 100개의 DNA 서열들에 대해서 약 25분이 걸린다. 개발된 시스템은 이러한 작업을 분석자의 간섭 없이 실행하고, 각 서열에 대해서 6가지 가능한 번역에서 가장 유사한 하나를 추천해 준다.

번역된 아미노산 서열들을 확인하기 위해서 분석자는 각 서열에 대해서 관련된 아미노산 정보가 주어진 NCBI BLAST를 실행할 것이다. BLAST에 대한 응답 시간은 시스템 부하에 달려 있다. 이러한 실험에 의하면 응답 시간은 약 15초 걸리고, 질의를 입력하고 결과를 복사하기 위한 조작 시간은 약 10초 걸린다. 따라서 100개의 항목에 대해서 약 41분이 걸리게 된다. 이와 같은 작업은 분석자의 간섭 없이 개발된 시스템으로도 완료될 수 있다.

위에서 보여준 과정에 대해서 전통적인 방법으로 데이터베이스와 도구들을 사용하면 분석자는 100개 항목에 대해서 약 1시간 31분을 소요하게 된다. 개발된 도구는 분석자의 간섭 없이 자율적인 방법으로 위와 같은 일을 수행하므로 분석자는 번거로운

작업에서 벗어날 수 있다.

5. 결론 및 향후 과제

생물정보학 도구의 역할은 시간 낭비와 번거로운 수작업을 자동으로 실행함으로서 분석 부담을 줄이고, 분석 작업에서 생산성을 증가하는 것이다. 많은 분자 생물학적 연구들은 알려진 데이터로부터 대량의 DNA 서열을 수집하고, 차후 연구들을 위해서 미리 필요한 아미노산 서열로 번역하는 것을 요구한다.

본 논문은 이와 같은 요구를 충족시키기 위해서 설계하고 구현된 시스템을 보여주었다. 개발된 시스템은 웹 로봇의 도움으로서 한 번에 DNA 서열들을 수집하고, 가능한 번역을 통한 가장 긴 ORF로 평가하여 가장 유사한 아미노산 서열들을 추천한다.

차후 연구들로는 Pattern Finder를 위한 정규 표현 질의 지원, Biostatistics Viewer를 위한 부가적인 생물통계학적 정보 제공과 같은 부가적인 서비스를 위한 기능성에 대해서 추가하는 것이 남아 있다. 본 논문에서는 로봇 에이전트의 도움으로 알려진 생물학적 데이터베이스로부터 질의 서열과 일치하는 다중 DNA 서열들을 수집하고, 주석 부분이나 제목 부분이 유전자 온톨로지 정보를 고려한 질의 문장과 함께 존재할 수 있는 것처럼 질의와 일치하는 서열을 선택하기 위한 기능의 구현에 대해서 연구하였다.

참 고 문 헌

[1] P. Baldi, S. Brunak, Bioinformatics : The Machine Learning Approach(2nd Ed.), The MIT Press, 2001.

[2] NCBI National Center for Biotechnology Information, http://www.ncb.nih.gov/.

[3] NCBI GenBank, http://www.ncbi.nih.gov/Genbank/index.html.

[4] ExPAsy, http://www.expasy.ch/tools/dna.html.

[5] UniProtKB/Swiss-Prot, http://www.ebi.ac.uk/swissprot/.

[6] Entrez, http://www.ncbi.nlm.nih.gov/Entrez.

[7] R. Durbin, S. R. Eddy, A. Krogh, G. Mitchison, Biological Sequence Analysis : Probabilistic models of protein and nucleic acids, The Cambridge University Press, 1998.

[8] S. Alurn, Handbook of Computational Molecular Biology(Eds.), Chapman & Hall/CRC, 2006.

[9] T. A. Brown, Genomes(2nd ed), Oxford, United Kingdom : Wiley-liss, 2002.

[10] A. D. Baxevanis, B.F. F. Ouellette, Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins(Eds.), John Wiley & Sons, Inc, 1998.